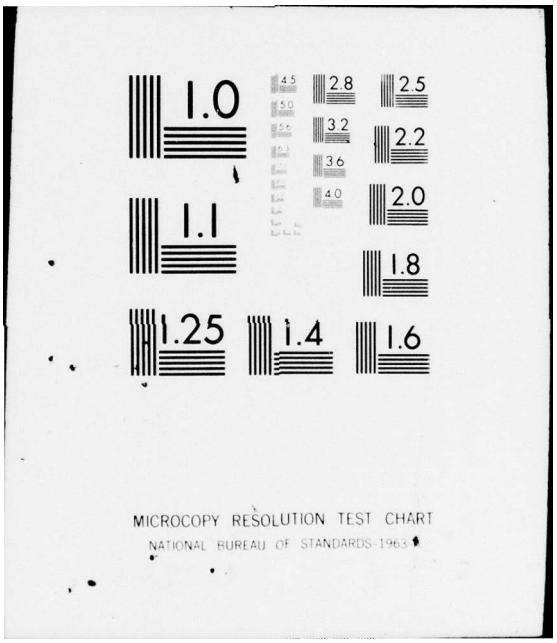
AD-A045 138 DESMATICS INC STATE COLLEGE PA F/G 12/1
IS RIDGE REGRESSION A PANACEA.(U)
SEP 77 D E SMITH N00014-75-C-1054
UNCLASSIFIED TR-106-5 NL

1 OF 1
AD
A045138

END
DATE
FILMED
11-77
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963

9 (12)

— STATISTICS —

— OPERATIONS RESEARCH —

— MATHEMATICS —

DESMATICS, INC.

P.O. Box 618
State College, Pa. 16801

DESMATICS, INC.

P. O. Box 618
State College, Pa. 16801
Phone: (814) 238-9621

Applied Research in Statistics - Mathematics - Operations Research

IS RIDGE REGRESSION A PANACEA.

by

Dennis E. Smith

TECHNICAL REPORT, NO. TR-106-5

September 1977

22 p.

This study was supported by the Office of Naval Research
under Contract No. N00014-75-C-1054, Task No. NR 042-334

391 156

TABLE OF CONTENTS

I. INTRODUCTION

Consider the multiple linear regression model

$$\underline{y} = \underline{X}\beta + \underline{\varepsilon}$$

where

$\underline{y}$ is an n x 1 vector of observations,

$\underline{X}$ is an n x p matrix of full rank,

β is a p x 1 vector of unknown parameters,

and $\underline{\varepsilon}$ is an n x 1 vector of errors.

In addition to the usual assumptions that $E(\underline{\varepsilon}) = \underline{0}$ and $\underline{V}(\underline{\varepsilon}) = \underline{I}\sigma^2$, it will also be assumed from the onset that $\underline{\varepsilon}$ is Normally distributed.

The most common estimator of β is the ordinary least squares (OLS) estimator $\hat{\beta}$ which is given by

$$\hat{\beta} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{y} .$$

Under the stated assumptions $\hat{\beta}$ has the following five properties:

(1) $\hat{\beta}$ minimizes $\underline{\varepsilon}'\underline{\varepsilon} = (\underline{y} - \underline{X}\beta)'(\underline{y} - \underline{X}\beta)$. That is, $\hat{\beta}$ is the value of β which minimizes the error sum of squares (the sum of the squared deviations of the observations from their expected values).

(2) $\hat{\beta}$ is the best linear unbiased estimator of β. That is, of all linear unbiased estimators, $\hat{\beta}$ is the one with minimum variance.

(3) $\hat{\underline{\beta}}$ is the maximum likelihood estimator of $\underline{\beta}$.

(4) $\hat{\underline{\beta}}$ is Normally distributed with mean $\underline{\beta}$ and covariance matrix $\sigma^2(\underline{X}'\underline{X})^{-1}$.

(5) The mean square error (MSE) of $\hat{\underline{\beta}}$ is

$$MSE(\hat{\underline{\beta}}) = E[(\hat{\underline{\beta}} - \underline{\beta})'(\hat{\underline{\beta}} - \underline{\beta})] = \sigma^2\Sigma(1/\lambda_i),$$

where the λ_i's are the eigenvalues of $\underline{X}'\underline{X}$.

It should be noted that if $\underline{X}'\underline{X}$ has any very small eigenvalues, $MSE(\hat{\underline{\beta}})$ will be extremely large. Thus, because $MSE(\hat{\underline{\beta}})$ is equivalent to the expected squared distance between $\hat{\underline{\beta}}$ and $\underline{\beta}$, the OLS estimator is likely to produce estimates which are quite far away from $\underline{\beta}$. Small eigenvalues will result, for example, when the $\underline{X}$ matrix is highly collinear, that is, when one column is close to being a linear combination of other columns.

II. RIDGE REGRESSION

With the objective of reducing MSE, Hoerl and Kennard [7,8] suggested the use of what they termed "ridge regression", so named because of its relationship to ridge analysis [3,6], a technique used in investigating fitted quadratic response functions. Ridge regression involves the use of an estimator which depends on the choice of a number $k \geq 0$. This ridge estimator, given by

$$\hat{\underline{\beta}}_k = (\underline{X}'\underline{X} + k\underline{I})^{-1}\underline{X}'\underline{y}$$

$$= \underline{W}_k\underline{X}'\underline{y} ,$$

has the following five properties:

(1) For $k = 0$, $\hat{\underline{\beta}}_k = \hat{\underline{\beta}}$. That is, the OLS estimator is a special case of the ridge estimator.

(2) For $k > 0$, $(\hat{\underline{\beta}}_k)'(\hat{\underline{\beta}}_k) < \hat{\underline{\beta}}'\hat{\underline{\beta}}$. That is, $\hat{\underline{\beta}}_k$ is shorter than $\hat{\underline{\beta}}$.

(3) $\hat{\underline{\beta}}_k$ is Normally distributed with mean $\underline{W}_k\underline{X}'\underline{X}\underline{\beta}$ and covariance matrix $\sigma^2\underline{W}_k\underline{X}'\underline{X}\underline{W}_k$. Thus, if $k \neq 0$, $\hat{\underline{\beta}}_k$ is a biased estimator since its mean is not equal to $\underline{\beta}$.

(4) The MSE of $\hat{\underline{\beta}}_k$ is $MSE(\hat{\underline{\beta}}_k) = \sigma^2\Sigma[\lambda_i/(\lambda_i + k)^2] + k^2\underline{\beta}'\underline{W}_k^2\underline{\beta}$. Because $\hat{\underline{\beta}}_k$ is a biased estimator, its MSE includes a bias term in addition to the variance term.

(5) There always exists a value $k > 0$ such that the ridge

estimator $\hat{\underline{\beta}}_k$ has smaller MSE than the OLS estimator $\hat{\underline{\beta}}$.

In other words, this last property states that although $\hat{\underline{\beta}}_k$ is biased for $k > 0$, there does exist a value of $k > 0$ such that the resulting bias is more than offset by a reduction in variance. Thus, the corresponding ridge estimator does provide a reduction in the MSE of the OLS estimator. The proof of this by Hoerl and Kennard [7] has been the basis of the excitement over the use of ridge regression. The statistical literature is now well-represented by articles discussing various aspects of ridge regression, for example [2, 4, 5, 9, 10, 11, 12, 13, 14, 15, 16, 17].

A. A SIMPLE EXAMPLE

For illustrative purposes, consider the problem of estimating the mean of a univariate Normal distribution. For this problem, the regression model $\underline{y} = \underline{X}\beta + \underline{\varepsilon}$ is, of course, appropriate. However, $\underline{\beta}$ is a 1×1 vector (i.e., a scalar β which denotes the population mean) and the $\underline{X}$ matrix is equal to $\underline{1}$, an $n \times 1$ column vectors of ones. Thus, the ridge estimator of the mean β is given by

$$\begin{aligned}\hat{\beta}_k &= (\underline{1}'\underline{1} + k\underline{I}_1)^{-1}\underline{1}'\underline{y} \\ &= (n + k)^{-1}\Sigma y_i \\ &= n\bar{y}/(n + k) .\end{aligned}$$

From property (4) on the preceding page,

$$MSE(\hat{\beta}_k) = n\sigma^2/(n + k)^2 + k^2\beta^2/(n + k)^2 .$$

By differentiating with respect to k, it can be verified that the value $k = \sigma^2/\beta^2$ yields the minimum MSE.

Unfortunately, the optimum value of k involves not only the unknown variance σ^2, but also the unknown parameter β which is to be estimated. This is also true for the general regression situation. Thus, the problem of how to choose the value of k must be considered.

B. CHOOSING THE VALUE OF k

In their original papers [7,8], Hoerl and Kennard discuss the use of a device they termed the "ridge trace", which is a plot of each component of $\hat{\underline{\beta}}_k$ against k. Their primary guideline is to choose a value of k where the system stabilizes, that is, where the components come together and more or less flatten out. However, relying on an "eyeball" judgement like this means that my choice of k and your choice of k may be completely different. In fact, your choice of k tomorrow may be different from your choice today, even for the same problem. Thus, there is no objective way to evaluate the performance of a ridge estimator chosen in this manner.

To overcome the objections to using a subjective estimate of k, a number of people have suggested various estimators for k. For example, Hoerl, Kennard, and Baldwin [9] proposed the estimator

$$k = p\hat{\sigma}^2/\hat{\underline{\beta}}'\hat{\underline{\beta}}.$$

MacDonald and Galarneau [13] suggested an estimator which has the "correct length." Since

$$E(\hat{\underline{\beta}}'\hat{\underline{\beta}}) = \underline{\beta}'\underline{\beta} + \sigma^2\Sigma(1/\lambda_i)$$

the quantity

$$Q = \hat{\underline{\beta}}'\hat{\underline{\beta}} - \hat{\sigma}^2\Sigma(1/\lambda_i)$$

is an unbiased estimator of $\underline{\beta}'\underline{\beta}$. Therefore, for $Q > 0$ MacDonald and Galarneau suggested using the estimator $\hat{\underline{\beta}}_k$, where $\hat{\underline{\beta}}_k$ is chosen such that

$$\hat{\underline{\beta}}_k'\hat{\underline{\beta}}_k = Q .$$

For $Q < 0$, they suggested choosing k equal to some prespecified constant k_0. Two specific choices are:

(1) $k_0 = 0$ (the OLS estimator)

(2) $k_0 = \infty$ $(\hat{\underline{\beta}}_k = \underline{0})$.

Both sets of authors carried out simulations to evaluate the performance of the proposed ridge estimators. In general, the ridge estimators did well in some portions of the parameter space and not too well in other portions.

It should be noted that when k is not a constant, the distributional and MSE properties previously listed for a ridge estimator do not hold, since these properties are conditional on k. Therefore, there is no guarantee that a value of k chosen by examination of the ridge trace or by application of some rule will yield an MSE which is smaller than that provided by the OLS estimator.

As an aside, it should be noted that in a Bayesian framework if $\underline{\beta}$

has a prior Normal distribution with mean $\underline{0}$ and covariance matrix $\sigma_0^2\underline{I}$, then the posterior mean of $\underline{\beta}$ has the same form as a ridge estimator with

$$k = \sigma^2/\sigma_0^2.$$

So, in a sense, by choosing k from the data rather than before the data is taken, ridge regression involves an a posteriori selection of the prior distribution.

C. TRANSFORMATIONS

With the usual regression model which includes a constant term, a problem may be considered in a number of different forms. For example, the original independent variables X_{ij}, centered independent variables $(X_{ij} - \bar{X}_{.j})$, or standardized independent variables

$$(X_{ij} - \bar{X}_{.j})/[\Sigma_i(X_{ij} - \bar{X}_{.j})^2]^{1/2}$$

may be used. In addition, the dependent variable may be centered or standardized.

In any event, all of these transformations result in the same least squares estimator $\hat{\underline{\beta}}$. Usually the standardized form is used in calculations, since it is less susceptible to round-off error. In this situation the $\underline{X}'\underline{X}$ matrix is given in correlation form. Most papers devoted to ridge regression have also assumed $\underline{X}'\underline{X}$ in correlation form, but this is not required by any of the underlying theory. However, unlike the least squares situation, ridge regression will produce different results for each transformation on the independent variables. That is, a ridge estimator <u>is not invariant</u> to the form of the model.

III. THE PURPOSE OF A REGRESSION INVESTIGATION

All regression investigations do not have the same purpose. Four possible purposes are:

(1) Estimation of $\underline{\beta}$

(2) Prediction of $\underline{y}$ (i.e., estimation of $\underline{X}_0\underline{\beta}$)

(3) Hypothesis Testing

(4) Optimization and Control.

For (4) the warning of Box [1] should be recalled. He pointed out that "... to find out what happens to a system when you interfere with it you have to interfere with it (not just passively observe it)." So, if at all possible, multicollinearity should be avoided by means of a well-designed experiment. Otherwise, the experimenter may be out on a limb. For (3) the experimenter is left floundering without adequate distribution theory if a ridge estimator is used instead of the OLS estimator.

It should be noted that even if the OLS estimator does not result in a very accurate estimate of $\underline{\beta}$ [purpose (1)], this does not necessarily mean that it will do badly in predicting $\underline{y}$ [purpose (2)], as the following example illustrates.

Figure 1 summarizes a regression problem discussed by Marquardt [10]. Because the regression model includes two independent variables but no constant term, a two-dimensional plot is adequate for displaying the problem. As can be seen from the figure, the variables X_1 and X_2 are highly correlated.

In general, an $\underline{X}'\underline{X}$ matrix may be expressed as

Model: $\underline{y} = \underline{X}\,\underline{\beta} + \underline{\epsilon}$

$$\underline{y} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \qquad \underline{X} = \sqrt{0.5} \begin{bmatrix} 0.6 & 0.8 \\ 0.8 & 0.6 \\ 1.0 & 1.0 \end{bmatrix} \qquad \underline{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

$$\underline{X}'\underline{X} = \begin{bmatrix} 1.000 & .980 \\ .980 & 1.000 \end{bmatrix} \qquad (\underline{X}'\underline{X})^{-1} = \begin{bmatrix} 25.253 & -24.747 \\ -24.747 & 25.253 \end{bmatrix}$$

$$V(\hat{\beta}_1) = V(\hat{\beta}_2) = 25.253\,\sigma^2$$

Figure 1: Marquardt's Regression Example

$$\underline{X}'\underline{X} = \sum_{1}^{p} \lambda_i \underline{v}_i \underline{v}_i'$$

where

$\lambda_1 > \ldots > \lambda_p > 0$ are eigenvalues of $\underline{X}'\underline{X}$

and

$\underline{v}_1, \ldots, \underline{v}_p$ are the corresponding normalized eigenvectors.

This representation of $\underline{X}'\underline{X}$ indicates how well the data space is covered, while a similar representation of $(\underline{X}'\underline{X})^{-1}$ indicates how well the parameter space is covered. As indicated in Figure 2, for this example 99% (1.98/2.00) of the variability in the data space is along the line $X_1 = X_2$, while $V(\hat{\beta}_1 + \hat{\beta}_2)$ is approximately 1% (0.51/50.51) of the total variance in the parameter space.

It should be noted that although $V(\hat{\beta}_1)$ and $V(\hat{\beta}_2)$ are large $(25.253\sigma^2)$, $V(\hat{y}_I) < .75\sigma^2$ for any predicted mean value within region I, the region defined by the data. Even some distance from this data region, things aren't too bad along the first principal component axis [for example, at point A $V(\hat{y}_A) = 1.01\sigma^2$], but aren't too good along the other axis [for example, at point B $V(\hat{y}_B) = 19.12\sigma^2$]. Points A and B are equidistant from the center of the data region I .

The moral of this is that if the concern is with predicting values of $\underline{y}$ within the region covered by the data, the results of using the OLS estimator are reasonable even if the estimate of $\underline{\beta}$ is not very accurate. In general, multicollinearity does not prevent good predictions of mean responses or of new observations, so long as these

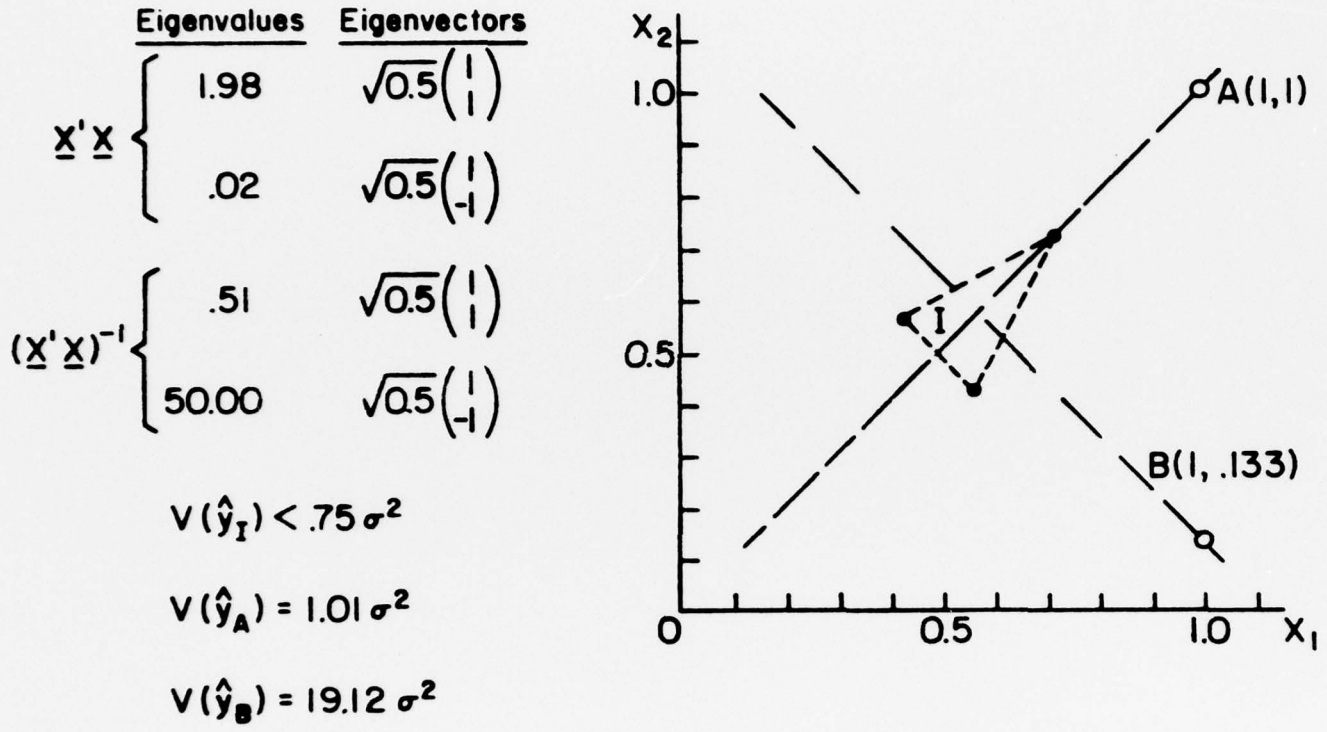


Figure 2: Coverage of the Data Space and Parameter Space

predictions are made for points within the data region. Of course, for points outside this region it must be realized that extrapolation may be dangerous regardless of whether an OLS estimator or a ridge estimator is used.

IV. SOME QUESTIONS

This report began with the question "Is ridge regression a panacea?" and will end with a number of other questions.

(1) Is $\underline{y} = \underline{X\beta} + \underline{\varepsilon}$ the true model or an approximation?

Most research establishing the usefulness of ridge estimators is based on the explicit or implicit assumption that $\underline{X\beta}$ is the true model, even outside the data region. In many cases this may not be a reasonable assumption.

(2) Is the OLS estimator deficient or is the data deficient?

There are two ways of viewing the results in a situation where there is a high degree of multicollinearity. The first is that the OLS estimator gives bad results. The second is that the data is inadequate for the estimation task at hand.

(3) Is there a structural relationship in the data?

If, for example, flow rate is always reduced as temperature is increased because of physical constraints, a high degree of collinearity will be present. In this type of situation, it might be well to incorporate this relationship directly into the statistical analysis.

(4) Can more data be obtained?

If predictions are to be made outside of the data region and if data can be observed there, the best bet may be to observe data there before attempting any estimation.

(5) Assuming that a ridge estimator may be useful sometimes, why not always use it?

It should be recalled that despite the degree of multicollinearity (even if there is complete orthogonality), there <u>always</u> exists a value of $k > 0$ such that $MSE(\hat{\underline{\beta}}_k) < MSE(\hat{\underline{\beta}})$.

(6) What about the lack of invariance?

If an experimenter does decide to use ridge regression, he or she must then consider the invariance problem, i.e., the form of the regression model to be used in actually carrying out the estimation process.

V. CONCLUSION

The potential user of ridge regression would be well-advised to consider seriously the questions listed in the previous section. In addition, he or she must remember that the use of a ridge estimator involves a trade-off between the chances of gains in certain regions of the parameter space and the chances of losses in certain other regions.

Although ridge regression may offer promise, its use as a routine analysis method is not without severe shortcomings. Therefore, the question in the title of this report is answered in the negative.

VI. REFERENCES

[1] Box, G. E. P. (1966) Use and abuse of regression. Technometrics, 8, 625-629.

[2] Conniffe, D. and Stone, J. (1973) A critical view of ridge regression. The Statistician, 22, 181-187.

[3] Draper, N. R. (1963) Ridge analysis of response surfaces. Technometrics, 5, 469-479.

[4] Guilkey, D. K., and Murphy, J. L. (1975) Directed ridge regression techniques in cases of multicollinearity. JASA, 70, 769-775.

[5] Hemmerle, William J. (1975) An explicit solution for generalized ridge regression. Technometrics, 17, 309-314.

[6] Hoerl, A. E. (1962) Application of ridge analysis to regression problems. Ch. Eng. Prog., 58, 54-59.

[7] Hoerl, A. E. and Kennard, R. W. (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 12, 55-67.

[8] Hoerl, A. E. and Kennard, R. W. (1970) Ridge regression: applications to nonorthogonal problems. Technometrics, 12, 69-82.

[9] Hoerl, A. E., Kennard, R. W., and Baldwin, K. F. (1975) Ridge regression: some simulations. Comm. Stat., 4, 105-123.

[10] Marquardt, D. W. (1970) Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics, 12, 591-612.

[11] Marquardt, D. W. and Snee, R. D. (1975) Ridge regression in practice. American Statistician, 29, 3-20.

[12] Mayer, L. S. and Willke, T. A. (1973) On biased estimation in linear models. Technometrics, 15, 497-505.

[13] McDonald, G. C. and Galarneau, D. I. (1975) A Monte-Carlo evaluation of some ridge-type estimators. JASA, 70, 407-416.

[14] Newhouse, J. P. and Oman, S. D. (1971) An evaluation of ridge estimators. RAND report No. R-716-PR.

[15] Obenchain, R. L. (1975) Ridge analysis following a preliminary test of a shrunken hypothesis. Technometrics, 17, 431-441.

[16] Sidik, S. M. (1975) Comaarison of some biased estimation methods (including ordinary subset regression) in the linear model. NASA Technical Note No. NASA-TN-D-7932.

[17] Thisted, R. A. (1976) Ridge regression, minimax estimation, and empirical Bayes methods. Division of Biostatistics, Stanford University, Technical Report No. 28.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 106-5	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* IS RIDGE REGRESSION A PANACEA?		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Dennis E. Smith		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-1054
9. PERFORMING ORGANIZATION NAME AND ADDRESS Desmatics, Inc. P. O. Box 618 State College, Pa. 16801		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 042-334
11. CONTROLLING OFFICE NAME AND ADDRESS Statistics and Probability Program (Code 436) Office of Naval Research Arlington, VA 22217		12. REPORT DATE September 1977 13. NUMBER OF PAGES 17
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Ridge regression
Ridge estimation
Multiple regression

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Consider the usual regression model $\underline{y} = \underline{X}\beta + \underline{\varepsilon}$ where $\underline{X}$ is a matrix of full rank, β is a vector of unknown parameters, and $\underline{\varepsilon}$ is a vector of random errors such that $E(\underline{\varepsilon}) = 0$ and $\text{Var}(\underline{\varepsilon}) = \sigma^2\underline{I}$. The procedure known as ridge regression has been offered as an alternative to ordinary least squares for estimating β, particularly in those situations where "severe" multicollinearity exists in $\underline{X}$. Ridge regression involves the use of a ridge estimator, which takes the

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

form $\hat{\beta}(k) = (X'X + kI)^{-1} X'y$ where $k \geq 0$.

The properties of ridge regression relative to those of ordinary least squares are discussed. Although ridge regression does appear to offer promise, its use as a routine analysis method is not without shortcomings. Therefore, the question in the title is answered in the negative.